Ameriflux Data Curation Portal Status and Learnings to Date

Berkeley Water Center Microsoft TCI
4/26/2006

Preamble

- Thanks for taking the time!
- This deck is a variant of an April status deck used internally in MSFT and with other e-science groups
- We're made some progress
- We're still in tool learning mode – many of the tools are new to us and our knowledge is the biggest stumbling block
- We're looking for feedback to improve and suggestions for next steps

2

Outline

- Ameriflux Data Overview (reminder)
- What does the "original" data look like?
- What schema are we using? Why?
- What portal workflows are we going to implement first?
- Current status – a few fun pictures
- Proposed next steps

3

Ameriflux Overview

149 Sites across the Americas

Each site reports a minimum of 22 common measurements.

Communal science – each principle investigator acts independently to prepare and publish data.

Data published to and archived at Oak Ridge.

http://public.ornl.gov/ameriflux/

Ameriflux Data Access Today

There are multiple (different) paths to the published data

Overview summary of (site:year) indicating data that has been submitted:

http://public.ornl.gov/ameriflux/viewstatus_Ameriflux.cfm

Single site .csv data download and simple graphics:

http://cdiac.esd.ornl.gov/programs/ameriflux/data_system/aamer.html

Multiple site .csv data download:

http://cdiac.esd.ornl.gov/programs/ameriflux/data_system/aallamer_pf.html

5

Ameriflux Data Access Today (fine print)

The single site page access includes only a subset of the sites indicated as having submitted data.

The multiple site download is a subset of the single site list.

A (hidden) ftp site holds older data (undocumented status) and on request data for specific investigators.

The available data is a mix of direct measurements and computation.

Data is subject to change at any time.

Data may differ from site to site. Some even include code!

Exact metrics differ from investigator to investigator. (Humidity may be .73 or 73.; Nulls and gaps may be null, zero, or -9999, or…)

Metadata is also dirty. (Eg reported latitude/longitude may vary across pages referencing site and within produced downloaded .csv file.)

The multi-site or multi-year downloads yield one big and likely unwieldy .csv file. (Can't import into excel, notepad, etc.) If the download actually completes.

Can view plots of one or two variables per site – little ability to compare sites or preview data available before download

6

Lesson #1: There are lots of slightly different “Ameriflux datasets”

The data published to the site is at least partially processed by the publishing investigator and is subject to temporal versioning.

Our Ameriflux Data Set

To simplify the problem, we’re using a partially curated set as our “original” data.

Derived from the larger “original” data by Mattias Falk.

Collected in self-identifying .csv files.

Headers are not “computational science clean”. Instead, we rely on the file naming convention to determine the site and investigator metadata.

8

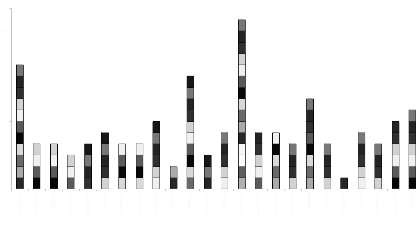
Ameriflux “Falk” Data Overview

24 sites with at least some data

No single year recorded at all sites

“Big” or ~150MB total

“Computer science dirty”
- Zero length files
- Not quite all units normalized
- Uneven directory structure



9

Ameriflux “Falk” Data Files

Comma separated file
- But commas are used in header

Directory encodes site
- But site can be aggregate

File name encodes site and year

Header includes site name, location, investigator, ecosystem type
- But not necessarily consistent with website

Data is every half hour or every hour

May not cover full year

Null data reported as -9999.

Sitename: Tonzi Ranch, California, USA

Location: Latitude: 38.4192 Longitude: -120.9503

Principal investigator: Dennis Baldocchi

Ecosystem type: Woody savannas

File creation date: 22JUL2005

Data Policy --Kindly inform the appropriate Principal Investigators of how you are using flux and site data and of any publication plans.

If the Principal Investigators feel that they should be acknowledged or offered participation as authors, they will let you know.

We assume that an agreement on such matters

will be reached prior to publishing and/or use of the data for publication. If your work directly competes with the Principal Investigators analysis, he or she may ask that they have the opportunity to submit a manuscript before you submit the one that uses their data.

In addition, when publishing a paper using flux data, please acknowledge the agency that supported the flux tower research.—

File Origin - This file was created at Oak Ridge National Laboratory by the AmeriFlux and FLUXNET data management groups.

These groups are supported by the U.S. Department of Energy and National Aeronautics and Space Administration.

This standardized file is identical in format to other standardized files provided here with a goal of aiding intersite comparisons, multi-site syntheses, and modeling activities.

Questions about these standardized files should be addressed to Tom Boden (bodenta@ornl.gov) FOR AMERIFLUX or Susan Holladay for FLUXNET.

FLUXNET draft as of 12/06/2004

YEAR, GAP, DTIME, DOY, HRMIN, UST, TA, WD, WS, NEE, FC, SC, H, SA, LE, SW, G1, TS1, TS1_DEPTH, TS2, TS2_DEPTH, PRECIP, RH, PA, CO2, VPD, SWC1, SWC2, RNET, PAR_in, RG_in, fRG_in, APAR, PAR_refl, RG_refl, RLONG_in, RLONG_out, H2O, E, Z1

YEAR, GAP, DTIME, DOY, HRMIN, m s-1, degrees C, degrees, m s-1, umol m-2 s-1, umol m-2 s-1, umol m-2 s-1, w m-2, w m-2, w m-2, w m-2, w m-2, degrees C, cm, degrees C, cm, mm, percent, kPa, umol mol-1, kPa, m3 m-3, m3 m-3, w m-2, umol m-2 s-1, w m-2, w m-2, umol m-2 s-1, umol m-2 s-1, w m-2, w m-2, w m-2, mmol mol-1, mg m-2 s-1, unitless

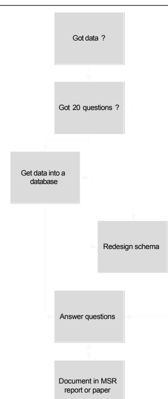
2004 ,-1 ,1.00000 , 1 ,000 ,0 ,7.03 ,8 ,0.33 ,-9999 ,-9999 ,-9999 ,-9999 ,-9999 ,-9999 ,-9999 ,-13.618 ,11.298 ,2 ,11.825 ,4 ,0 ,97.2 ,100.1 ,445.623 ,0.032 ,0.003 ,0.003 ,-47.769 ,0 ,-1.988 ,-9999 ,-9999 ,-0.248 ,-9999 ,-9999 ,-9999 ,-9999 ,-9999 ,-99992004 ,0 ,1.02083 , 1 ,030 ,0.398 ,8.48 ,16 ,0.22 ,-9999 ,-4.821 ,-9999 ,3.913 ,-9999 ,3.127 ,-9999 ,-10.481 ,11.486 ,2 ,11.903 ,4 ,0 ,97.7 ,100.1 ,422.674 ,0.029 ,0.003 ,0.003 ,-39.711 ,0.249 ,-1.59 ,-9999 ,-9999 ,0 ,-9999 ,-9999 ,-9999 ,-9999 ,-9999 ,-9999

10

Ameriflux Prototype

Demonstration vehicle for all parties
- Scientists ask “how does this help me?”
- Microsoft people ask “what does e-science really mean?”
- I ask “how do I use these tools?”

Following a modified Grayfomatics™ methodology
- Get a data set in house
- Develop 20 questions
- *Identify key workflow(s)*
- Load a database
- *Enable the workflow(s)*
- Answer the questions
- Iterate



(Italics are modifications for Ameriflux)

11

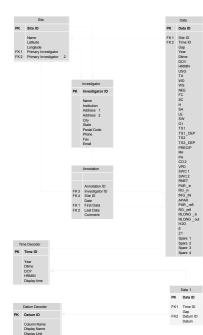
Initial Ameriflux Database Schema

Simple Data vector table
- Maps directly to csv file format
- Simple load
- Not expandable or self identifying
- May add Data1 table

Site, Investigator, and Annotation tables hold useful references
- Not really GIS
- Can we leverage STS or other common code?

Time Decoder table for MDX
- MDX time is motivated by business reporting; our time is motivated by diurnal and seasonal variations

Datum Decoder to drive displays
- Goal is to leverage database and avoid writing code



12

What was wrong?

- All text fields should be unicode
  - SQL Server Integration Services (SSIS) default is unicode; anything else requires more complex logic to convert
- No way of identifying a dataset across time
  - Our "original" data is actually processed by investigators who can republish the "same" dataset.
- No built-in ability to derive a working dataset for scientific use
  - Original data is gap filled, recalibrated, as well as used for graphs and statistics
  - Ability to version is core to the analysis – is this calibration algorithm better/worse than another?
- Data vector structure leads to accumulating code that knows about that structure
  - Not as reusable for other similar data or, more importantly, when federating two very different types of data

13

Lesson #2: Original data is in the eye of the beholder

One investigator's "raw" or "original" data is another's processed data.
So, after initial load of "original" data, derive a working dataset for scientific use.

Lesson #3: Abstracted datasets are your friends

Abstracted (normalized, pivoted) datasets should allow subsequent analysis tools to be common. In other words, interesting time series and statistical analyses can be written once and not per datum type.

Lesson #4: Keep the data load as simple as possible.

While the documentation explains how to extend SSIS with amazing scripting and complex flows, simple flows are easier to debug when the data is dirty. Even if you load into a dedicated database. Oh, and don't scrimp on space – keep everything. Also, add "superfluous" columns to track anything you might wonder about later.

Lesson #4 We need a clear data usage model

There are lots of twisty passages that computer scientist can go down. Which helps you more?

Datasets

- Datasets are collections of data used for a specific analysis or the result of measurement or simulation or library research or other information production activity
- All datasets have a unique identifier.
- Datasets may be *published* or *private*.
- Datasets are stored in an *original* or vectorized format or a *working* or atomized format.

18

Published Datasets Usage

- Published datasets:
  - Are discoverable by and usable by other investigators, although some access policy may apply.
  - Are curated in an archive
  - Have contents that are fixed (unchangeable)
- All datasets used to derive a dataset must be published when that dataset is published.
  - The unique identifiers are used to derive a published dataset are published with the published dataset.
  - The unique identifers can be used to establish provenance over time and subsequent use by other scientists.
- Published datasets may be:
  - *Published* by uploading to an archiving portal
  - *Downloaded* from an archiving or other caching portal
  - <Open question> *Deleted* when there is no other published dataset that has been derived from the dataset.

19

Private Dataset Usage

- Private datasets:
  - Are private to an individual investigator or small group of investigators.
  - May change at any time.
  - Before publishing, the dataset unique identifier can is used only by an individual scientist or small collaborating group of scientists.
- Private dataset operations:
  - *Import* is the ingest of an ad hoc shared dataset.
    - Analogous to download, but no discovery, access control or guarantee of stability.
  - *Export* is the emission of an ad hoc shared dataset.
    - Analogous to publishing, but no discovery, access control or guarantee of stability.
  - *Creation* can result from actual measurements, simulation, literature research or any other means.
  - *Derivation* is the result of some sort of data transformation (eg gap fill or calibration) on one or more existing published or private datasets.
  - *Appending* adds data to an existing unpublished dataset.
  - *Pruning* removes data from an existing unpublished dataset.
  - *Deletion* eliminates a non-published dataset.

20

Working or Atomized Datasets

- Normalized schema:
  - Allows a wide variety of datum types to be mixed in a single dataset
  - Enables analysis code and UI driving that code to be independent of datum type and private data formats.
  - Is what an experienced database programmer would use!
  - Not intuitive to most scientific programmers.
- Atomization requires:
  - All values must be interpreted indirectly through datum identifiers.
  - A specific datum identifier includes the name and units of the datum.
  - The data format is rigidly determined by the datum identifier.
    - For example, the datum identifier used to report humidity as ".73" differs from that used to report humidity as "73."
    - A clearinghouse for datum identifiers is necessary. Several are evolving. We simply assume there is one.
    - <Open question> how is a scientist going to be able to do that mapping simply and robustly given the richness?

21

Original or Vectorized Datasets

- Original datasets have private vectorized table format that can be mapped to a CSV file.
  - CSV files are the lingua franca today for data interchange and simple analysis via Excel or other tools
  - The column layout and data representation (eg units and scaling) are often private to investigators
  - Collaborations often develop a common format for dataset publishing.
  - Accommodating these vector formats are important to the scientist's workflow – especially since the data are often dirty and must be interpreted by the scientist.
- For each vector format, there is:
  - A private data table format. That format must include a dataset unique identifer.
  - Code that can import/export the vectorized data from/to a csv file. This includes a subset of the columns.
  - Code that can convert the private table rows to/from atomized data.

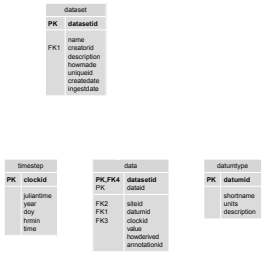
22

New Ameriflux Database Schema

23

Working Dataset

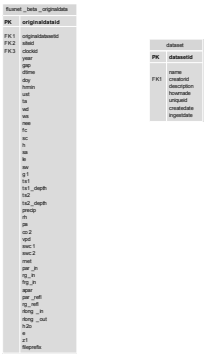
- Working datasets characterized by a name, creating investigator, description, and a GUID
- Actual data characterized by:
  - Dataset
  - Site (where collected)
  - Datum (for interpretatin
  - Clock (time)
  - Value
  - How derived still TBD, but definitely includes "gap"
- Datum characterization eventually should correspond to GIS/HIS or other public schema
- (Recall that the Ameriflux data are published on regular time intervals)



24

Original Dataset

- Original datasets characterized by a name, creating investigator, description, how constructed and a GUID
  - How constructed text still TBD, but includes “measure”, “simulate”, ???
- Data vector follows excel spreadsheet columns
  - Simple loading is a good thing
  - Additional column for fileprefix used to generate siteid key after load
  - clockid key generated after load from year/day/hrmin
  - Dtime unused, but retained until we know it’s unnecessary



Sites and Investigators

- Investigator loaded from processed scraping of http://public.ornl.gov/ameriflux/participants.cfm
  - Only some actually own towers
- Sites loaded from processed scraping of http://cdiac.esd.ornl.gov/programs/ameriflux/data_system/aamer.html
- A better long term solution would be to crawl each site page and cross reference investigators and other site metadata
  - What happens over time as investigators (particularly tower owners) come and go?
- Site:investigator determined after load of site (by parsing investigatortext and comparing with lastname

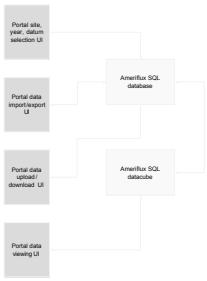
Annotations

- Unstructured comments by an investigator on any other table row
  - Logbook-like entries
  - Data inconsistencies, etc notes
  - Changes in sites, investigators
- Not intended as an data acquisition event log (eg low battery of a sensor)
- Research agenda is to observe how these are used and how they should be curated

annotation	
PK,FK1,FK2	**datasetid**
PK,FK4	investigatorId
PK	annotationid
	date
	comment
FK3	datumid
FK1	dataid
FK6	siteid
FK7	originaldataid

Portal Dataset Access

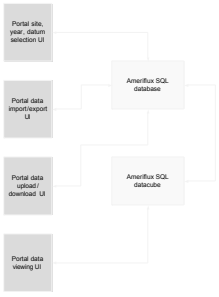
- The portal is the access point to the published dataset archive
  - Behind the portal is one or more archiving databases that contain published original datasets, published working datasets, and the investigator, site, datum, and annotation metadata that describe those datasets.
  - The actual implementation may be arbitrarily complex. The portal may query other database on behalf of the user and/or federate over many databases.
- “Interesting” published original or working datasets may be downloaded for private analysis
  - The downloaded data can be exported as a CSV vector or ingested into an analysis database which uses the same schema as the portal
  - If downloaded to an analysis database, that database may be co-located with the portal, on the scientist desktop, or any other database server location.



(follows SkyServer – adds annotation and plans for federation)

Data Analyses

- Common data analyses are supported by any analysis database or datacube
  - The same UI can be used on the portal or on the desktop
  - At a minimum, the portal must support the common browsing analyses useful for determining whether a working dataset is “interesting”
- Analyses we’re thinking about for the prototype
  - Graph time series of selected variables over sites and times
  - Correlate selected variables over sites, times, and lagtimes
  - Compute simple statistics (means, standard deviations)
  - Tile plot diurnal and seasonal variations of a variable, product, or other statistic of multiple variables. We intend to fix the definition of “season” for simplicity.



(follows SkyServer – less need for spatial analyses, more need for time series)

Prototype Workflows

- Ingest portal metadata
  - Includes site, investigator, timestep datum definitions
- Publish original dataset
  - We focus on the “Falk” dataset; but this can generalize to others including other datasets on the Ameriflux web site.
- Publish working dataset
  - Our datum types follow the “Falk” dataset and we make no attempt to use a standard data definition
- Select “interesting” original dataset(s) and download original dataset(s)
  - Select by site or time span
  - Full vector always downloaded without further interpretation
- Select “interesting” original or working dataset(s) and download working dataset(s)
  - Select by site or time span or datumIDs
  - Selecting specific datumIDs cause subset of working dataset to be downloaded
- Select “interesting” original or working dataset(s) and derive a working dataset
  - Select by site, time span or measurement
- Select “interesting” working dataset(s) and make another working set
  - Subset by site, time span, or datum
  - Apply specific processing (eg gap fill)
- Use the working dataset
  - Graph time series of selected variables over sites and times
  - Autocorrelate selected variables over sites, times, and lagtimes
  - Compute simple statistics (means, standard deviations)
  - Tile plot diurnal and seasonal variations of a variable, product, or other statistic of multiple variables.

Current Status

- Visual Studio project leveraging SSIS for metadata and data loading of site, investigator, timestep and datum tables working from spreadsheets.
  - Dynamic update of spreadsheet from web site not currently considered.
- Raw data load project working from "Falk" spreadsheets.
  - All "original data" loaded
  - Working data set generated with air temperature (TA), CO2 flux (FC), CO2 in canopy (SC), and precipitation (PRECIP) for development
  - Working data set is hourly only (temporary hack)
- Met with potential contractor for portal download and data selection web service. Early sketches of UI developed jointly. Lawyers, lawyers, lawyers.
- Talking with CUASHI HIS HODM team (David Tarboton, Reza Wahaji, Blair Jennings, Ilya Zaslavsky, et al) on common schema longer term.

31

Lesson #5: IT groups are like scientists.

The contractor's "customer model" for the code to be written should follow their practices when working with a medium-size business. The IT groups at such companies want to be able to debug and/or make minor changes in deployed applications without having to (re)call a contractor. That translates to being able to drag/drop or simple script or maybe even simple SQL, but no heavy programming. Note the similarities with how much a "scientist" wants to know about .NET, C#, etc.

Lesson #6: Our documentation sucks.

The tutorials (on line and in the documentation) are great for getting started, but don't give that many hints and/or assume a developer mind set or are evangelizing a technology. A non-expert (me) can browse that and learn all sorts of things but it's hit or miss.

For a good time, compare the online SSIS tutorial method for iterating over files to the SSIS book method. Is one better than the other? One is certainly –easier-.

Lesson #7: (In progress) If you drag and drop to program, how do you debug?

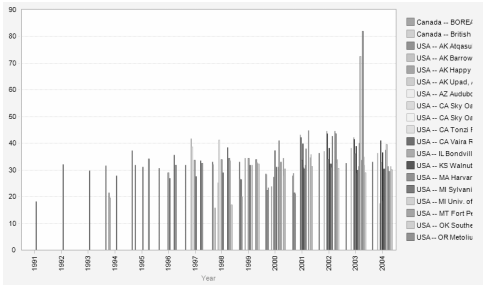
I've now managed to get breakpoints established and even a grid view going. But I've had best luck with a variant of the printf – adding a nvarchar(MAX) column to a table and stuffing suspicious text in along the way.

Now for some fun: What's wrong with the following pictures?

All of following slides were generated off the portal on my laptop.

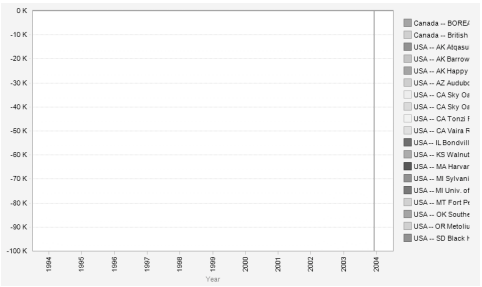
DISCLAIMER:
We're still very new to the tools and learning how to do things.

Maximum Air Temperature



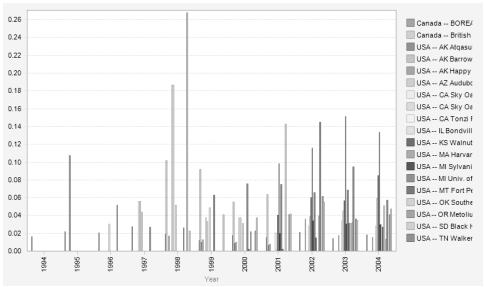
36

Minimum Precipitation



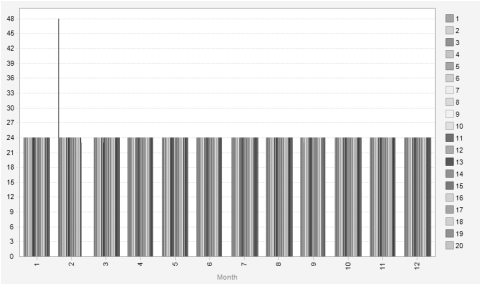
37

Mean Precipitation (w/o Vaira Ranch)



38

Count of Air Temperature measurements per month at Tonzi Ranch



39

Next Steps

- Learn just enough MDX/SQL voodoo to enable your next questions (samples from Gretchen):
  - Scatter plot temperature from Vaira and Tonzi Ranch
  - How frequently did it rain in 2003 at Sky Oaks?
  - How many days a year was the CO2 flux greater than average?
  - How frequently did it rain in a given year?
  - How many days in a given year had an above average carbon flux (fc) but a below average soil water content (swc)?
  - When the temperature dropped more than 2 degrees in a one hour period (two consecutive measurements), what was the probability that rain would fall in the next two hours?
  - ?????
- Make the LBL portal available for your own data browsing.
- Get the selection download UI working.
- Continue discussions with the CUASHI HIS HODM folks.

40

Microsoft

Your potential. Our passion.